SCIENCE CHINA
Life Sciences

SPECIAL TOPIC: Lifeomics and translational medicine
• REVIEW •

March 2013 Vol.56 No.3: 220–227
doi: 10.1007/s11427-013-4440-1

Information engineering infrastructure for life sciences and its implementation in China

ZHU WeiMin[1,2,3*], ZHU YunPing[3,4] & YANG XiaoLing[1]

[1]*Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & School of Basic Medicine, Peking Union Medical College, Beijing 100730, China;*
[2]*Taicang Institute of Life Sciences Information, Suzhou 215400, China;*
[3]*National Center of Protein Sciences (Beijing), Beijing 102206, China;*
[4]*Beijing Proteome Research Center, Beijing 102206, China*

Received December 18, 2012; accepted December 31, 2012

Biological data, represented by the data from omics platforms, are accumulating exponentially. As some other data-intensive scientific disciplines such as high-energy physics, climatology, meteorology, geology, geography and environmental sciences, modern life sciences have entered the information-rich era, the era of the 4th paradigm. The creation of Chinese information engineering infrastructure for pan-omics studies (CIEIPOS) has been long overdue as part of national scientific infrastructure, in accelerating the further development of Chinese life sciences, and translating rich data into knowledge and medical applications. By gathering facts of current status of international and Chinese bioinformatics communities in collecting, managing and utilizing biological data, the essay stresses the significance and urgency to create a 'data hub' in CIEIPOS, discusses challenges and possible solutions to integrate, query and visualize these data. Another important component of CIEIPOS, which is not part of traditional biological data centers such as NCBI and EBI, is omics informatics. Mass spectroscopy platform was taken as an example to illustrate the complexity of omics informatics. Its heavy dependency on computational power is highlighted. The demand for such power in omics studies is argued as the fundamental function to meet for CIEIPOS. Implementation outlook of CIEIPOS in hardware and network is discussed.

biological database services, omics informatics, information engineering infrastructure for pan-omics studies

Citation: Zhu W M, Zhu Y P, Yang X L. Information engineering infrastructure for life sciences and its implementation in China. Sci China Life Sci, 2013, 56: 220−227, doi: 10.1007/s11427-013-4440-1

1 Omics studies require large-scale information engineering infrastructure

1.1 Omics studies are instrumental tools for modern life sciences

Thanks to the revolutionary breakthrough of genome sequencing technology, and rapid advance of the technologies of gene chips, mass spectroscopy and next generation sequencing (NGS), a series of omics platforms featuring holistic study of biological molecules have been developed in the last decade or so. Some mature platforms, such as genomics, transcriptomics and proteomics, have transformed traditional biological research approach from examining bio-molecules in isolation to a new paradigm by studying biological functions of genes, transcripts and proteins into complex molecular networks. An -ome is such a network that, in the form of map or profile, embodies the interactive relationships between these molecules. Cross-omics studies that capture the vertical relationships between different layers of networks have become important tools to investigate natural state of biological systems. As such 'holistic' ap-

*Corresponding author (email: wmzhuworld@gmail.com)

proach finds wider applications in modern life sciences, a group of new platforms, such as metabolomics, interactomics, physiomics and phenomics, has been added into repertoire of omics studies. As a result, a rich collection of omics data types has been generated. With the cost of omics experiments rapidly reduced, and the throughput of omics technology platforms steadily increased, the volume of omics data has been explored. These characteristics of omics data in types and volume, further complicated by their complex relationships and diversity of biological systems, have brought us at the early stage of omics studies an unprecedented informatics challenge in data collection, storage, processing, analysis, distribution and application [1].

1.2 Omics studies are the driving force for modern life sciences to enter the era of the 4th paradigm

Informatics challenges brought by omics studies were coined as 'grand challenge' five years ago in information sciences. The 4th paradigm, a data-intensive scientific discovery, evolved from earlier paradigms of discovery methodologies: experimental science, theoretical science and computer science [2,3], is the solution to the grand challenge. Some typical examples of grand challenge in life sciences include the construction of evolution tree of life, drug design and discovery, decoding the mystery of biological systems and the search for the cure of cancers. Natural increase of computation power governed by Moor's law no longer meets the computational demands these challenges bring. Possible solutions to the challenges can only be sought by the close collaborations between disciplines of natural sciences, mathematics and computer sciences, and by integration of different informatics solutions in software tools and hardware technologies, such as high performance computing (HPC), paralleled storage and fast network. Information engineering has long been an integral part of traditional data-intensive disciplines, such as high-energy physics, meteorology, geology, geography, and environmental sciences. To life science community, however, its importance has never been fully recognized until the advent of omics studies.

1.3 The international landscape of state-level informatics infrastructure for life sciences

In order to fully take advantage of rich omics data, and establish their leading roles in life science research, a number of industrial countries have launched respective initiatives to build national information infrastructure. Cyber Infrastructure Task Force for Grand Challenge, spearheaded by National Science Foundation (NSF), proposed in 2009 to create National Cyber-Infrastructure in USA (http://www.nsf.gov/dir/index.jsp?org=OCI). Life sciences were included as one of the important applications. The task force consists of 6 working groups, in areas of computational methods & algorithms, software, HPC, data & visualization, training and the scientific applications of grand challenge, featuring intimate collaboration between the components of 'software' and 'hardware' of the cyber infrastructure. As early as the 90's of the 20th century, Department of Energy (DoE) of USA funded the first international human genome database (GDB). In 2009, the Office of Biological and Environmental Research at DoE initiated another important information engineering project—System Biology Knowledgebase (Kbase), to collect, curate, annotate, integrate and mine omics data of plants, microbes and environmental samples, providing mechanisms to model and simulate biological systems, generate new scientific hypotheses to guide the design of new experiment (http://kbase.science.energy.gov/). German BioEconomy Council directly links national biomedical information infrastructure with the immense economical potential the modern life sciences could bring, provides long-term funding for the integration of scientific research and information engineering technologies (http://www.europabio.org/industrial/news/bio-economy-council-first-report-german-government/). Major mandate of ELIXIR project [4], initiated by European Bioinformatics Institute (EBI) and European Union, participated by 14 member states in Europe, is to create pan-European bio-information infrastructure to support biomedical researches, and to translate rich biomedical data into applications of medicine, environment, biotechnology, and social sciences. Australia, although a late-comer in traditional biological research, has made major financial commitment by government in the creation of Australian National Data Service (ANDS), and built HPC-Avoca with the peak computing power at PFlops level, the most powerful HPC designated for life sciences at the time of writing (http://www.vlsci.org.au/). Avoca is mandated to provide huge computing power for disease diagnosis and treatment, drug discovery, and investigation into some major diseases such as cancer and epilepsy.

1.4 The current state of Chinese information engineering infrastructure for life sciences

To welcome the paradigm shift, many Chinese information scientists from data-intensive disciplines, including those from life sciences, organized a workshop in 2011. A proposal was produced to bolster China's informatics research for data-intensive scientific domains. It calls for a stable state funding, and the mechanism to share data. It is worth noting also that it signifies the importance to train cross-disciplinary talents who have the knowledge and skills in both scientific domain and information technology. The proposal is the first step to the right direction calling for creation of national digital infrastructure. It lacks, however, careful analysis of details in depth for the planning and implementation of such an infrastructure. Chinese omics scientists, represented by those in proteomics study from Beijing Proteome Research Center (BPRC), and those in genomics study from Beijing Genomics Institute (BGI), have

made tremendous progress in their respective researches, produced works respected by their colleagues world-wide, and contributed to the fact that China has led the world in data production. However, China has not 'cash out' the dividend matched to its investment into biological researches. Because it does not have a strong informatics infrastructure, those raw omics data either left at hard drive untouched, or exported without deeper analysis and exploration, not to mention to realize their application potentials in knowledge mining and medicine. Therefore, there is a great urgency to create such an infrastructure. The information engineering infrastructure for pan-omics studies (CIEIPOS) should not only meet the basic needs of data-intensive discovery of modern life sciences, but also have its strategic importance for China to prepare for its biomedical renaissance in the next 5 to 10 years, and to become a leading nation in modern life sciences.

2 Bioinformatics resources

2.1 International bioinformatics centers and their limitation in serving Chinese life scientists

Thirty years ago, National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) and European Bioinformatics Institute (EBI, http://www.ebi.ac.uk/) started to collect nucleotide sequences and created GenBank and EMBL databases respectively. Today, their rich collection of databases has accumulated to petabytes (PB) level of biological information, including data types such as nucleotide and protein sequences, protein signature/domain & 3D structure, molecular interaction & network, and chemicals and pathway. In the last few years, a deluge of omics data is produced from various omics platforms and ever-increasing number of personal-omics studies. The volume of these types of data has exceeded the totality of those curated data deposited at NCBI and EBI in the last 30 years. Together, they are the records of knowledge about lives, the treasure for us to share, to explore, and to be applied for the wellbeing of mankind. However, these data were collected, categorized, and distributed in/from USA and UK. The service portfolios were designed towards the users geographically adjacent to them. As stated in the previous section, Chinese scientists have produced great amount of biological data and China has become a major data contributor to international biological repositories. Because of the lack of comprehensive local capacity of 'software' and 'hardware' in processing, analyzing and managing these big data, China has become the biggest raw data exporter. Ironically, when it comes to retrieve these datasets for reanalysis, Chinese scientists often find it difficult, either because of restricted bulk downloading by hosting sources, or the constraint of network bandwidth. A local copy of comprehensive collection of public biological data, well curated with rich annotations and data types, should thus be the first target of CIEIPOS. This component, a data collection, integration& query system (CIQS), constitutes a series of bioinformatics services like those provided by NCBI or EBI in meeting the basic needs of biological researches in China.

2.2 Status of Chinese bioinformatics services

In the last decade, Chinese bioinformaticians have made unremitting efforts in realizing the goal of CIQS. The Center for Bioinformatics (CBI, http://www.cbi.pku.edu.cn/) at Beijing University is well-known for its comprehensive list of database mirrors, stable and regularly updated, of popular international databases. It is a highly appreciated service by Chinese scientists for quicker downloads of these complete databases. CBI also champions in developing a set of novel and popular biological databases, such as PlantTFDB for plant transcription factors, SynDB for protein set related to synapse or synaptic activity, and PathLocDB for sub-cellular locations of 40000+ metabolic pathways. Software tools, such as a generic genome browser ABrowser, bioinformatics resource management and configuration platform Weblab, and a popular protein-coding potential calculator CPC, are also among their innovative developments. Shanghai Center for Bioinformation Technology (SCBIT, http://www.scbit.org:8080/pages/index.do) has been focusing its efforts on the categorization of public nucleotide sequence and expression data into a group of database services, with added value of manual curation and annotation. Its Hotdata database aggregates supplementary datasets of journal articles, providing biologists with a unique data resource. Biologic Medicine Information Center (BMICC, http://www.bmicc.cn/web/share/home/), affiliated with Chinese Medical Academy of Sciences, offers statistical data analysis services on a set of valuable Chinese physiological and psychological survey data. It also classifies and serves a group of logically integrated biological databases, produced by Chinese laboratories such as ProteomeView of proteome from BPRC (http://proteomeview.hupo.org.cn/) and NONCODE of ncRNA [5] jointly developed by Institutes of Biophysics and Computing Technology at Chinese Academy of Sciences. BMICC is mandated to build translational data chain from biology, physiology to medicine. Refer to an excellent review article by Wei et al. [6] for more bioinformatics services offered by other Chinese groups. Useful as they are, these public Chinese bioinformatics services, however, have not become indispensable tools for Chinese life scientists to conduct their researches, due to their 'boutique workshop'-like operations: scattered resources, low engineering quality & stability, simple repetition in contents, and poor usability. In the area of capacity to store and manage data, organizations in China are either coping or at a primitive stage, with the only possible exception of BGI.

Despite continuous appeals and persistent efforts over the years by practitioners, this state of resource-scattering and lack-of-momentum in Chinese biomedical services has not

been fundamentally changed. One of the main causes, besides weak leadership and severe lack of centralized funding, is that informatics has long been considered to play an auxiliary role in the realm of traditional biology. The importance to manage and reuse data has never been fully recognized. As a result, professionals in 'making database and software' are marginalized by the performance assessment and promotion system, cross-disciplinary talents with both domain knowledge and informatics know-how are extremely difficult to attract and retain in practice. China is currently lagging behind the major western countries by 15–20 years in the application of information technology to life sciences. In the era of paradigm change to data-intensive discovery and exponential accumulation of omics data, if we remain inactive in eradicating such unjustified bias from the processes of decision-making, administration, and academic recognition, China will face serious consequence of missing this rare opportunity and to be left behind forever.

Big biomedical data have great challenges of their own when it comes to serve them to users. These challenges originate from complexity in data types, relationships and dimensions, heterogeneity in format, syntax and semantics, dynamics and stochasticity in contents and nature, compounded by the massiveness in volume. The challenges to integrate, search and visualize them have been unprecedented. So, implementation of CIQS at CIEIPOS requires close collaborations between scientific disciplines, synergies between science and technology, and, more importantly, integration of 'software' (including human factors) and 'hardware' components in information technology, all are indispensable guaranty of success.

3 Data integration, retrieval and visualization at CIEIPOS

Biological data, large in volume and complex in types and relationships, bring us great informatics challenges in integration, retrieval and visualization, in which integration is the problem at its most acute.

A wide spectrum of integration approaches have been proposed by information scientists in the last 20 years. Some examples include data modeling, either loosely-coupled views or tightly-coupled data warehouses, ontology-driven semantic integration, XML (extensive markup language)-based syntactic technique, and database-links that provide the mechanism to link a database entry to those in other databases. A few of these *modi oprandi* have been implemented by bioinformaticians, such as Sequence Retrieval System (SRS) [7], a database link-based tool developed at EBI, a data model-based system caBIG [8] and data warehouse-based system featuring reversed star schema BioMart [9]. These tools have been instrumental in providing biologists with 'one-stop-shopping' service. However, the general consensus reached through their practice and by user communities is that no single integration technology is sufficient to address the challenge of biomedical data integration. The design principle of data integration of CIQS, therefore, should be driven by use cases, considering data characteristics inherited in data size, degree of complexity and heterogeneity, and update frequency. The relationship between datasets is another important parameter to be taken into consideration. An integration framework should be constructed by using arsenal of approaches, from the modeling of well-structured data, semantic mapping & depiction between datasets from independent sources, the syntactic interoperability at the application layer, to explicit cross-reference embedded in database entries. Domain standards in data formats, procedures and data quality should be the key part of the framework to ensure the data interoperability. Due to the dynamic nature of biology, the integration framework should also be flexible in managing data type on-demand and keep its architecture scalable to accommodate different size of datasets. The construction of this sophisticated information system is an endeavor which could only be accomplished by close collaboration between information specialists and domain scientists.

One of prerequisites for such a system to be successful is unambiguous delineation of the relationships between datasets, so that sophisticated questions can be asked across databases and against multi-layers of -omes, so that an integrated view of multi-facets of a biological system could be returned. The challenges to build such a system lie in subtle balance between search performance and accuracy, and an effective yet friendly interface to navigate search results. Relational Database Management System (RDBMS) does not furnish such a balance when it comes to big data, neither the view-based federated query system [10], nor the metadata-driven system [11]. View-based system is capable of complex cross-database searches. But it suffers severely at performance when datasets are big, not to mention the overhead in maintaining mappings between data model and datasets. Apache Lucene technology [12] improves search performance by flexibly configured and pre-computed indices. It is weak at semantics, however, so tedious tuning rules and ranking scheme are required to make it useful. Entrez by NCBI (http://www.ncbi.nlm.nih.gov/Entrez/) and EBeye by EBI (http://www.ebi.ac.uk/ebisearch/) are today's most popular biological search engines. They also use pre-computed indices for quick data retrieval. Database cross-references, often carefully curated and embedded in data entries, are the main mechanism to link current data entry to relevant records in external databases. Entrez and EBeye share a similar feel-&-look interface and browsing functions, both having a good balance of search performance and accuracy. The major issue of both search engines is the lack of a top level navigation mechanism for the ease of browsing. They require users to have a certain degree of familiarity with databases, in format and content, since databases are often in different layouts, which are especially

true for EBI databases, and individual database record/entry is the entry point to view data details. In addition, there is no global filter to shorten hit lists. Filters usually exist only at individual database level. Both NCBI and EBI are continuously working at improving functions and usability of their search engines. There are obvious obstacles, however, if changes do not touch the root of problems, which lie deep at their underneath database architecture and operational model. Taicang Institute for Life Sciences Information (TILSI) recently launched its cross-database search engine Bioso! (www.bioso.org/). At the early implementation of its design, it considers the balance of performance and specificity in answering queries, offering an encyclopedic view to integrate search results in chapters (biomedical themes) and sections (information classes of a theme), with ongoing efforts in semantic integration via such approaches as database-links, ontology and metadata. This 'one-stop-shopping' feature is also supplemented by traditional hit lists of individual databases, for which filters are provided to narrow the hits.

Data visualization has become an irreplaceable tool to intuitively view and comprehend information and the relationships borne by complex biological and omics data of molecules, processes and systems. It is an integral part of CIQS at CIEIPOS to view complex data, to interact with experimentalists in high-throughput platforms and to assist in the generation of new scientific hypothesis. Low-dimensional presentation techniques, such as those in list and tabular formats, are no longer sufficient to achieve these objectives, which opens up new horizons for bioinformaticians to present data richer in display and higher in dimensions [13]. Visualization tools developed so far can be categorized into three different classes: (i) Static graph, drawn on a set of pre-computed data. Examples include Ensembl genome browser visualizing chromosomal and gene structure at adjustable resolutions; ProVit [14] aggregates and visualizes binary relationship of protein-protein interactions into complex network; KEGG database [15] provides a series of manually illustrated metabolic pathways and cellular biological processes, visually demonstrates complicated molecular interactions in the context of biochemical processes, and heatmap, dot/profile plots are common forms to illustrate expression data at transcription and translation levels. (ii) Three-dimensional (3D) model, plotted in 3D format, also from pre-computed data. A good example is the illustration of PDB data into 3D protein structure by tools such as RasMol, Jmol and SWISS-PDBView. (iii) Visualized simulation, static graphs or 3D models dynamically changed by the input of different parameters. CellDesigner [16] simulates metabolic pathway through modeling of biochemical events. There have been many useful visualization tools developed in the last decade. CIEIPOS should take advantage of these public resources and collaborates closely with computational biologists and graphic artists to develop more sophisticated, powerful and easy-to-use visualization tool to meet even-increasing demands from biomedical research and omics studies.

4 CIEIPOS' informatics services to omics studies

4.1 Traditional bioinformatics centers & their collaboration model with omics consortia

Omics data, due to their large size and proximity to experimental platforms, requires a unique management model. In the last 4–5 years, due to a deluge of omics data, NCBI and EBI have been seriously challenged in how to collect, manage and present these data. Short read archive (SRA) [17] is the answer, proposed by International Nucleotide Sequence Database Collaboration (INSDC) and implemented jointly by its member databases Genbank, ENA and DDBJ at NCBI, EBI and DDBJ. It specifically addresses an urgent need of an international central deposition for NGS data. Unlike EBI and NCBI's traditional model that focuses on value-added curation and annotations for submitted analyzed data, SRA takes NGS unassembled data, but leaves data dissimulation, interpretation and integration to users and, more importantly, to international collaboration bodies and consortia of genomics studies. ENCODE project [18] collects genome data, processed or unprocessed, from 7 American laboratories, centrally managed and presented at Stanford University. Data interoperability and comparability are ensured by a set of standards ranging from data format, quality control in data collection and processing, to experimental designs. 1000 genome project [19], participated by 8 countries, aims to compile genetic variations of human genomes in different races. It employs standards and data processing procedures similar to those of ENCODE. The data are deposited into project repositories at NCBI and EBI, either of which provides synchronized releases of variation datasets. Ensembl [20] and UCSC [21] databases collect and integrate genome data from variety of sources, including those from SRA, offering reference genome data along with rich annotations. The model that permanent repositories for metadata and raw data are provided by traditional bioinformatics centers such as NCBI and EBI, while international omics collaborations and consortia take on the role of data standardization and dissimulation, such as analysis, annotation and integration of data, has been widely adopted by today's bioinformatics community. The other early adopter of this model is proteomics community. PRIDE is an EBI database [22] archiving submissions of peptide/protein identification from proteomics experiments, with no or minimal curation and annotation activities. Peptidome [23] is a similar proteomics database at NCBI, closed down as a result of a funding cut. Tranche was designed as the permanent home for raw mass spectroscopy data, now replaced by a similar raw repository at EBI. Standardization, annotation and integration of these data have been taken up largely by

initiatives and consortia under Human Proteome Organization (HUPO), such as Human Proteome Plan (HPP), Proteome Standardization Initiatives (PSI), and proteome knowledgebase such as PeptideAtlas [24]. Another example of the adoption of this model is transcriptomics community in its collection, management and integration of RNAseq data.

4.2 Capacity in processing and managing omics data must be an integral part of CIEIPOS

Omics data are the core pieces of information in modern life sciences. As the national center of biological information, the ability to sufficiently process and manage these data has to be an integral part of CIEIPOS's capacity, besides its CIQS component. However, omics data are usually produced locally by omics centers, such as those of genomics or proteomics in China. In addition, scientists at wet labs often have the need to interactively process and analyze data in near-real time when they are conducting experiments, which further signifies physical separation of omics data from CIEIPOS. A software system, consisted of data standards, software tools and workflow, has to be implemented at these centers, in order to ensure the integrability and quality of information. It is an important role for CIEIPOS to play in coordinating and facilitating the implementation of such a system by providing guidelines and design principles. CIEIPOS's another important role is to provide annotation tools and resources, by well-maintained databases and distributed annotation system (DAS) [25] or cloud computing environment, to infer functional and physiochemical features from known sources to omics data. For more dispersed data sources, CIEIPOS has the responsibility in liaisoning them with national omics center/consortium in order to ensure their needs are met in data standards and data analysis. Omics centers shall provide CIEIPOS with the mechanism to obtain metadata, the access to raw data, and the data transportation technique necessary for CIEIPOS to conduct deep analysis, integration and simulation. An intimate collaboration between CIEIPOS and omics centers is fundamental instrument for China to support omics studies at national level, whereas the feasible implementation of such collaboration is a prerequisite to materializing omics informatics support at CIEIPOS. A feasibility study project of the National Bioinformatics Center, funded by Academician Bureau of Chinese Academy of Sciences, was recently launched. The search for an optimal model of such collaboration will be an important part of the study.

4.3 Omics data informatics has huge demand for computational power

Proteomics is a holistic study of species of proteins in a biological system and their relationships. A proteome is the central hub to collate biological functions between omes at various layers. The following paragraph will outline the complexity in managing and analyzing mass spectrometric (MS) data, so as to illustrate the great demand for computational power by omics informatics. A mass spectrometer has annual data production rate at 10 terabytes, averaging at 27 gigabytes a day. A modern national proteomics center, such as Phoenix Proteomics Research Infrastructure under construction, is equipped with several dozens of such instruments, with daily data output averaging at 1 terabyte. A typical workflow of MS data analysis is consisted of many steps based largely on *brute force* computation: spectral peak extraction, peptide mapping, protein inference and association of functional and physiochemical features with the identified proteins. Peptide mapping is an analytic process to query predicted sequences computed from spectra against one or more reference sequence libraries/databases. Computationally, this process is known as a NP-complete (NPC) problem, a problem with the uncertainty in finding a solution for, and greedy at CPU time. UniProtKB [26] is a typical example of reference database used for peptide mapping, with 30.31 million peptide/protein sequences, and 9.7 trillions of letters (release 2013_02, http://www.ebi.ac.uk/uniprot/Documentation/). If quantitative measurement, post-translational modification and alternative splicing are also considered, together with spectrum–sequence mapping, the computational requirements will be increased exponentially in many folds. Moreover, unlike genome, different organs/tissues at different developmental states/environments express different proteomes. It determines high dimensional nature of proteomics data and our ability to analyze proteomics data is seriously constrained by the 'curse of dimensionality', in addition to NPC/NP-hard problems as discussed above, as far as computation is concerned. Dimension-reduction is a technique often employed to approximate the problem to non-deterministic polynomial or polynomial ones, a powerful computational environment is nevertheless a fundamental requirement of CIEIPOS.

Modelling and simulation of omics data span an enormous temporal or spatial range [27]. For instance, a spatial range of e^{15} could be reached if data analyses span from the level of molecules at a resolution of 100 Å (metabolome, genome, transcriptome and proteome), to reactome and localizome at the resolution of 1000 Å, physiome at 1–10 micron, and finally to the human living environment measured at kilometers. For temporal range, from a molecular interaction occurring at femtoseconds (fms), formation of protein structure in seconds, cellular processes in minutes, hours, and days, to the evolution events occurring in millions of years, it can reach the scale of e^{30}. Pheonome and Physiome are assemblies of multi-omics data by modeling and simulation, starting vertically from genomes and transcriptomes. They offer systematic view of bio-physiological processes, such as regulation process with biological effects of activation and suppression, signal transduction in biochemical metabolism, chemical and cellular transportation

events, cell division and the growth of a tissue. The modeling of such complex bio-physiological processes requires not only huge computation power to accomplish, but also a large local storage space, high performance and parallel processing architecture. At Victorian Life Sciences Computation Initiative (VLSCI), HPC-Avoca has been used to model rhinovirus, scientists are able to observe, at the first time, intimately of drug interaction on virus. Avoca has also been used to model cellular events for other organisms, and simulations in diagnosis and treatment of diseases such as glaucoma, diabetics and cancers. In China, there is no such a HPC, designated to life sciences, equivalent to the power of Avoca yet. Most of modeling and simulation studies are still confined to the 3rd paradigm of discovery with small data set and narrower temporal-spatial range.

5 Outlook of hardware & network implementation at CIEIPOS

As illustrated above, CIEIPOS requires a robust data management and processing system to dissect data of various sources and a powerful integration system to discover complex relationships to transform facets of biological information into a body of knowledge. These systems are only possible with the availability of infrastructural support of powerful computational resources. China has been equipped with the technologies to build powerful HPCs of commercial quality. Local vendors have also had the ability to produce distributed storage to satisfy high I/O need of biological data. Infiniband at the speed over 40 GB promises high speed and high throughput communication between computation nodes and storage. Hadoop technology [28,29] offers a scalable and distributed software framework for HPC, with data processing power up to PB level. It speeds up computation by replicating data at multiple nodes, offers automatic error correction function by task trackers, and by the mechanisms to predict task execution and to execute a task in multiple nodes. These are the solid hardware and system software foundation for implementation of HPC environment at CIEIPOS. For network connectivity, Chinese Education and Research Network (CERNET) and Chinese Science and Technology Network (CSTNET) have been catching up in support of big data. Data providers, however, are still constrained by the subscription cost for sufficient bandwidth. Connections between CERNET and CSTNET are yet to be optimized and so is their connection with other public networks. These compound factors contribute to unstable operation of data providers in China and put great constraint on data flows between them and external data centers, as well as the access to their data resources by their users. For data consumers, it is also very difficult to obtain large datasets, due to the 'last foot' constraint. There are no easy solutions to these issues in foreseeable future. It might well be a plausible working solution, however, for CIEIPOS to escalate this urgent requirement of biomedical community to the level of national strategic planning and collaborate closely with network service suppliers, especially with CERNET and CSTNET, to make the case that, just like other data-intensive sciences, modern life science is one of their important applications. IPv6 is a new generation network protocol freely available to the public. Chinese government has introduced a series of policies to encourage commercial explorations of the technology, while Internet2, a science & technology network in USA, is also promoting the use of IPv6 network. It could be the future ultimate solution at CIEIPOS.

6 Conclusion

CIEIPOS is a long overdue infrastructure required for a renaissance of life sciences in China. The 21st century is known as the century of life sciences. To put it more precisely, it is the century of modern life sciences characterized by the exploration of omics information. In the first 10 years of this century, Chinese scientists have made considerable progress in biological research. Tremendous amount of data have been accumulated as the result. Because we lag behind at the key infrastructure to manage, analyze and utilize them, the gap between the need and the availability of such an infrastructure has hindered further advance of life sciences in the country. This infrastructural deficiency has also become impediment for China to translate research results into knowledge and medical applications. Meanwhile, a great deal of experiences in data management and processing has been accumulated in USA and Europe over the last 20 years. The gap to these two parts of the world, though existing, is not as wide as that occurring in China. Today, as the great potentials of omics studies have become more evident, these countries have already started to invest heavily in both 'hardware' and 'software', as well as in the human factor through training and retaining, in order to fill the gap. In conclusion, the creation of CIEIPOS is the first key step to enhance China's research power in modern life sciences. Just by seizing this rare opportunity would China succeed before it eventually takes the lead in fierce competitions worldwide to explore the potentials promised by pan-omics studies.

The work of Bioso! is not possible without generous financial support of Taicang government, Suzhou, China. We also thank Mr. Luo DengHui for his diligent editorial assistance to the manuscript.

1 Schadt E E, Linderman M D, Sorenson J, et al. Computational solutions to large-scale data management and analysis. Nat Rev Genet, 2010, 11: 647–657
2 Smith A, Balazinska M, Baru C, et al. Biology and data-intensive scientific discovery in the beginning of the 21st century. OMICS, 2011, 15: 209–212
3 Kolker E, Stewart E, Ozdemir V. Opportunities and challenges for

the life sciences community. OMICS, 2012, 16: 136–147
4 Crosswell L, Thornton J. ELIXIR: a distributed infrastructure for European biological data. Trends Biotechnol, 2012, 30: 241–242
5 Bu D C, Yu K T, Sun S L, et al. NONCODE v3.0: integrative annotation of long noncoding RNAs. Nucleic Acids Res, 2012, 40: D210–D215
6 Wei L P, Yu J. Bioinformatics in China: a personal perspective. PLoS Comput Biol, 2008, 4: e1000020
7 Zdobnov E M, Lopez R, Apweiler R, et al. The EBI SRS server—recent developments. Bioinformatics, 2002, 18: 368–373
8 Saltz J H, Oster S, Hastings S L, et al. Integrating heterogeneous rules-engine technologies with caGrid. AMIA Annu Symp Proc, 2007, 11: 1099
9 Smedley D, Haider S, Ballester B, et al. BioMart—biological queries made easy. BMC Genomics, 2009, 14: 22
10 Livne O E, Schultz N D, Narus S P. Federated querying architecture with clinical & translational health IT application. J Med Syst, 2011, 35: 1211–1224
11 van Vlymen J, de Lusignan S. A system of metadata to control the process of query, aggregating, cleaning and analysing large datasets of primary care data. Inform Prim Care, 2005, 13: 281–291
12 Shah P K, Perez-Iratxeta C, Bork P, et al. Information extraction from full text scientific articles: where are the keywords? BMC Bioinformatics, 2003, 4: 20
13 Gehlenborg N, O'Donoghue S I, Baliga N S, et al. Visualization of omics data for systems biology. Nat Meth, 2010, 7: S56–S68
14 Iragne F, Nikolski M, Mathieu B, et al. ProViz: protein interaction visualization and exploration. Bioinformatics, 2005, 21: 272–274
15 Zhou T T. Computational reconstruction of metabolic networks from KEGG. Methods Mol Biol, 2013, 930: 235–249
16 Funahashi A, Matsuoka Y, Jouraku A, et al. CellDesigner 3.5: a versatile modeling tool for biochemical networks. Proc IEEE, 2008, 96: 1254–1265
17 Leinonen R, Akhtar R, Birney E, et al. Improvements to services at the European Nucleotide Archive. Nucleic Acids Res, 2010, 38: D39–D45
18 ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature, 2012, 489: 57–74
19 Kuehn B M. 1000 Genomes Project finds substantial genetic variation among populations. JAMA, 2012, 308: 2322–2325
20 Flicek P, Ahmed I, Amode M R, et al. Ensembl 2013. Nucleic Acids Res, 2013, 41: D48–55
21 Meyer L R, Zweig A S, Hinrichs A S, et al. The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res, 2013, 41: D64–69
22 Vizcaíno J A, Côté R G, Csordas A, et al. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res, 2012, doi: 10.1093/nar/gks1262
23 Ji L, Barrett T, Ayanbule O, et al. NCBI Peptidome: a new repository for mass spectrometry proteomics data. Nucleic Acids Res, 2010, 38: D731–D735
24 Vizcaíno J A, Foster J M, Martens L. Proteomics data repositories: providing a safe haven for your data and acting as a springboard for further research. J Proteomics, 2010, 73: 2136–2146
25 Dowell R D, Jokerst R M, Day A, et al. The distributed annotation system. BMC Bioinformatics, 2001, 2: 7
26 Boeckmann B, Bairoch A, Apweiler R, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res, 2003, 31: 365–370
27 Hassanien A E, Milanova M, Smolinski T, et al. Computational intelligence in solving bioinformatics problems: reviews, perspectives, and challenges. Comp Intel in Biomed & Bioinform, SCI, 2008, 151: 3–47
28 Taylor R C. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics, 2010, 11: S1
29 Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. In: Proceedings of the 6th Symposium on OSDI, San Francisco, USA, 2004. 137–150